

Университет ИТМО

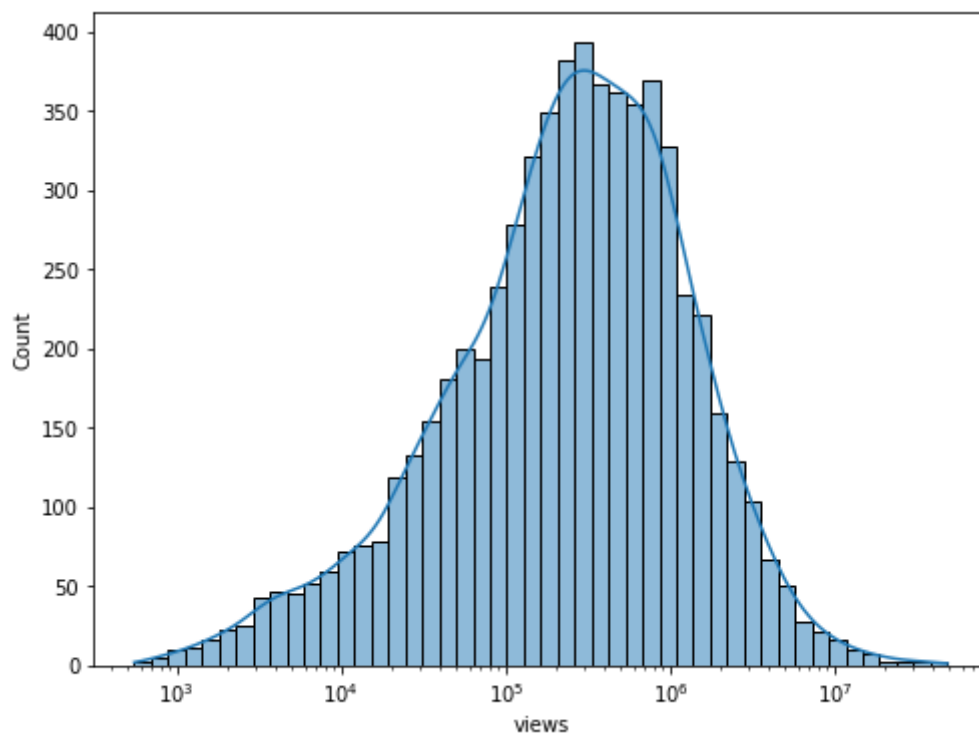
Практическая работа №4
по дисциплине «Визуализация и моделирование»

Автор: Сулейманов Руслан Имранович
Поток: 1.1 **Группа:** к3241 **Факультет:** ИКТ
Преподаватель: Чернышева А.В.

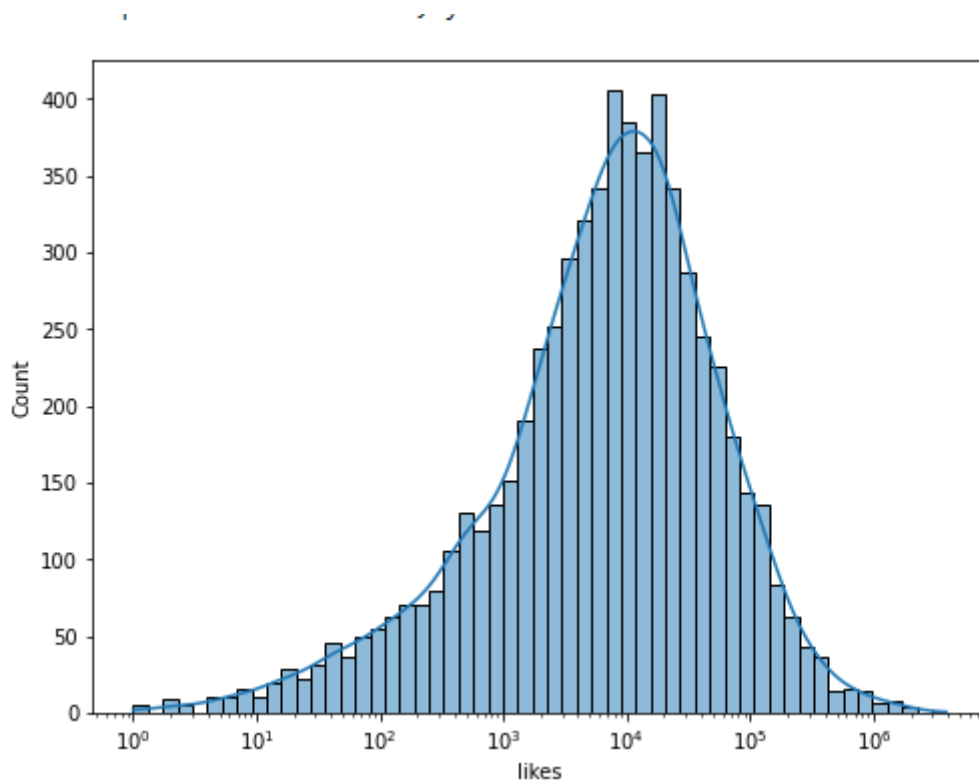
Санкт-Петербург, 2021 г.

Для начала, повторим все графики, так как мне повезло с датасетом, новых результатов быть не должно:

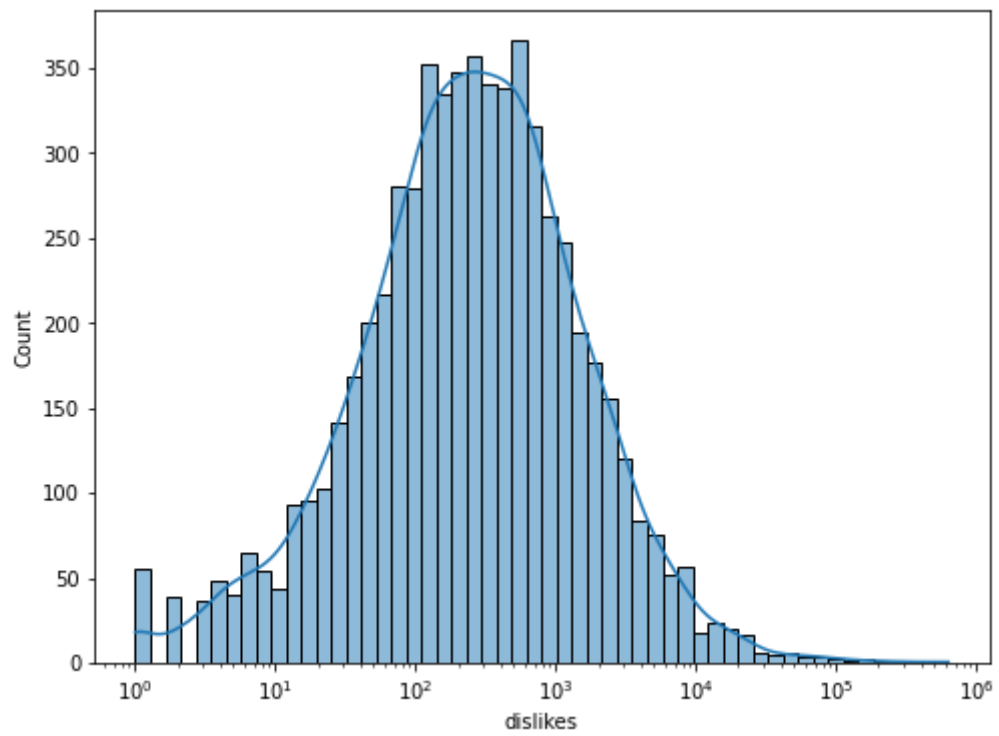
1. Количество видео с одинаковым количеством их просмотров



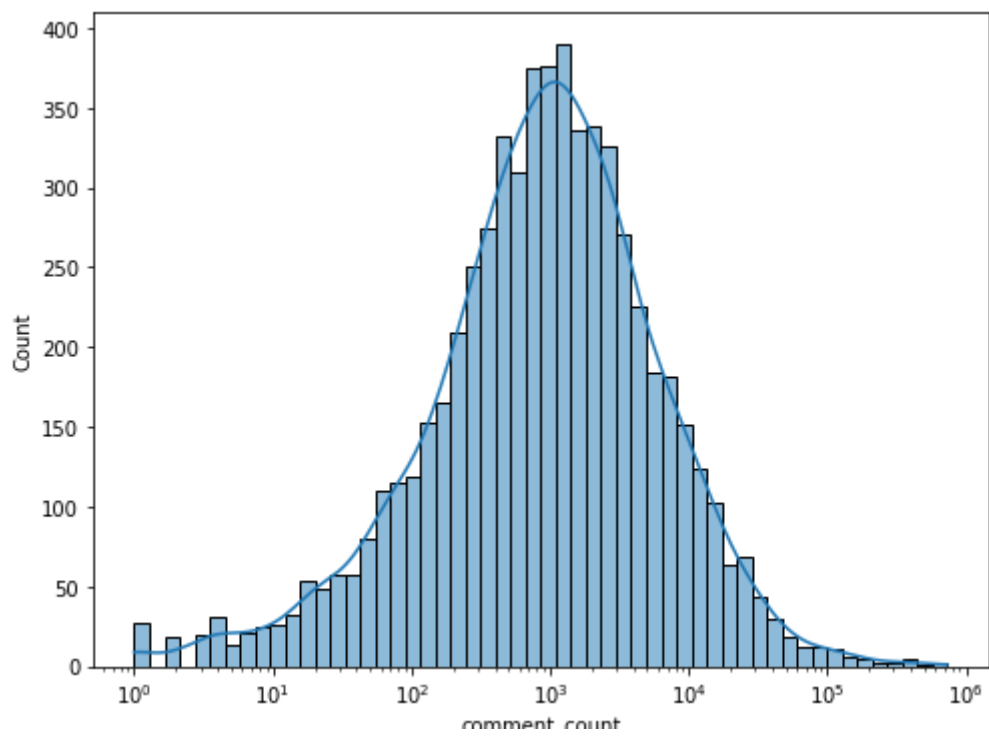
2. Количество видео с одинаковым количеством лайков



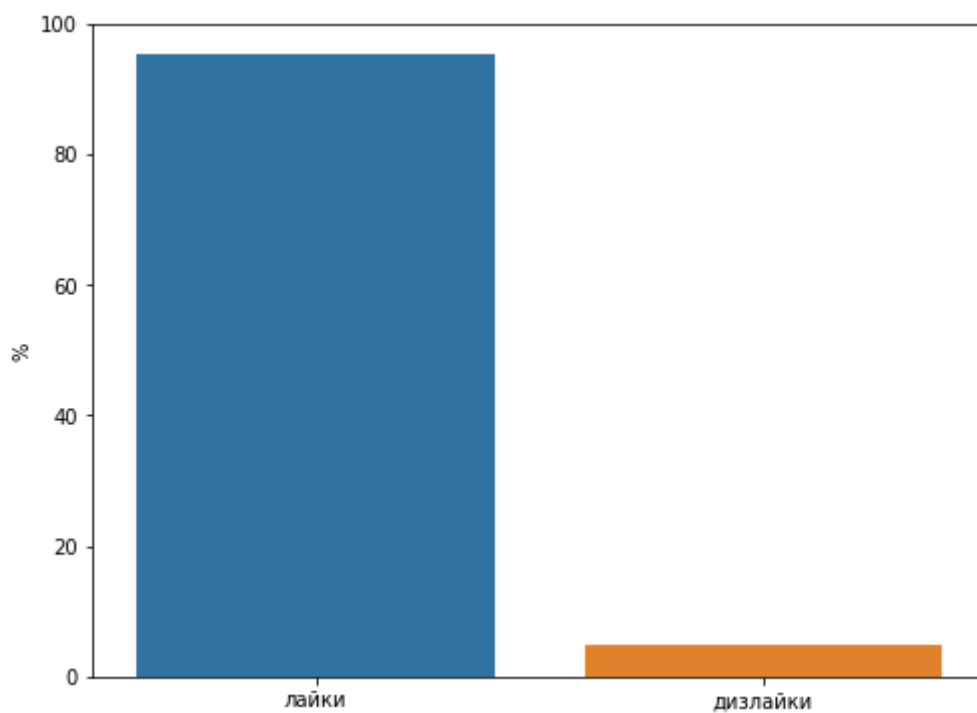
3. Количество видео с одинаковым количеством дизлайков



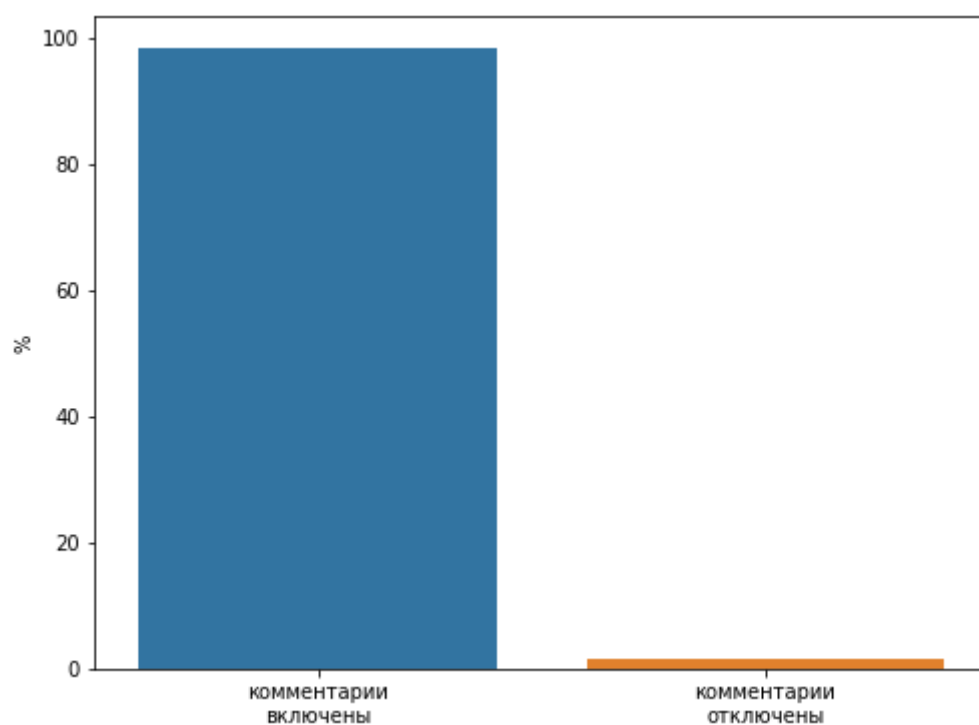
4. Количество видео с одинаковым количеством комментариев



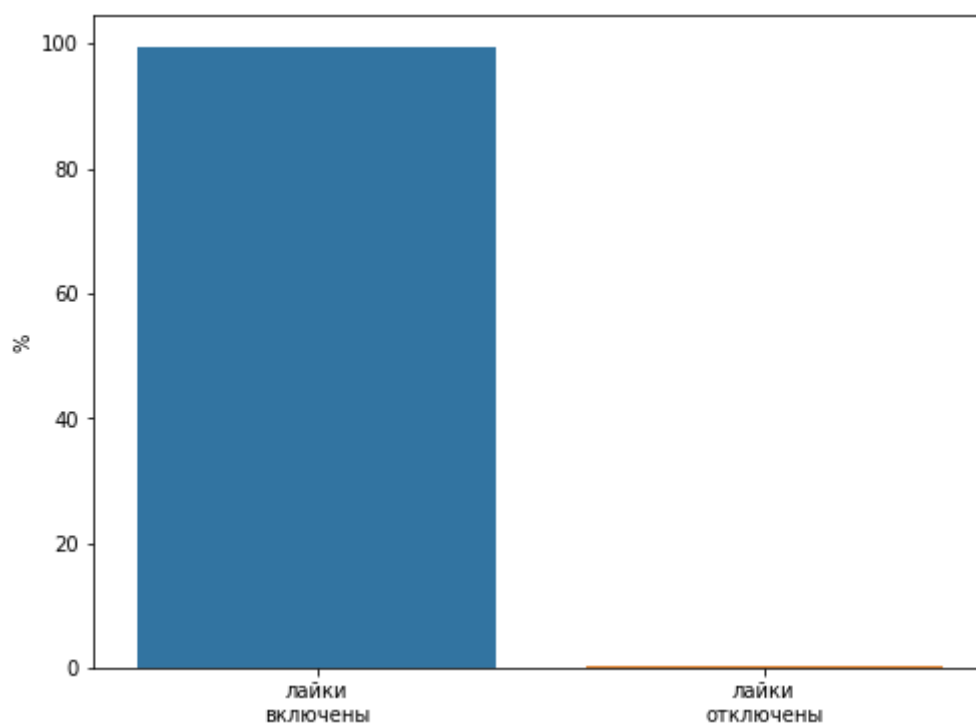
5. Соотношение лайков и дизлайков



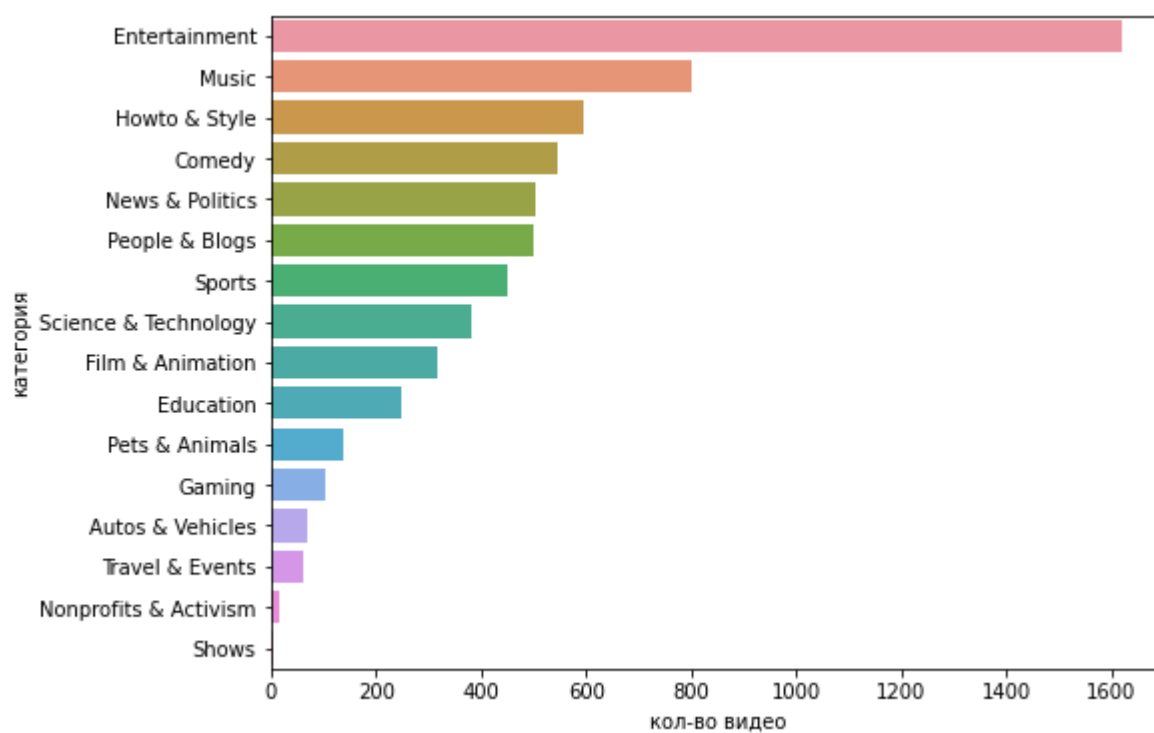
6. Процент видео с отключенными комментариями



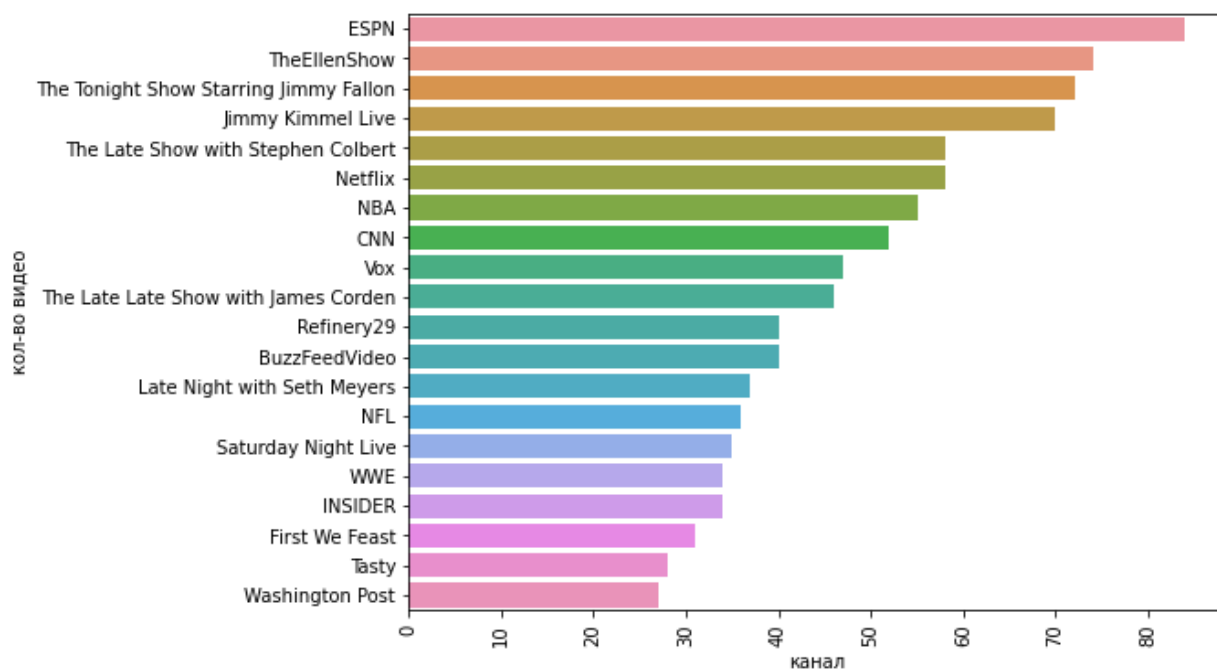
7. Процент видео с отключенными лайками



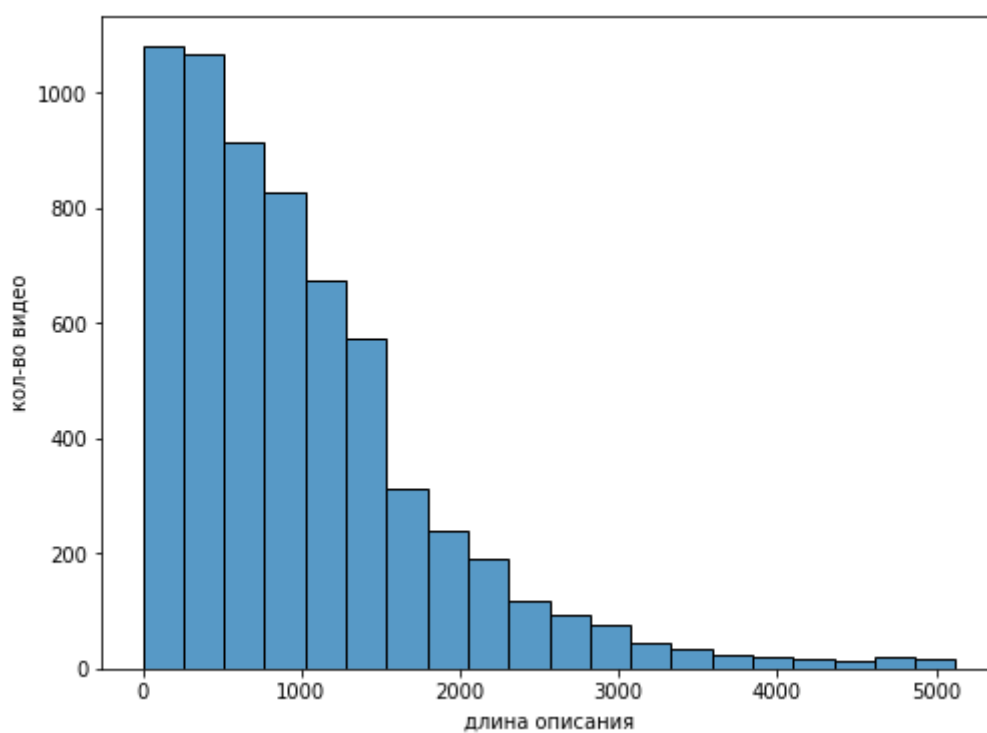
8. Топ популярных категорий



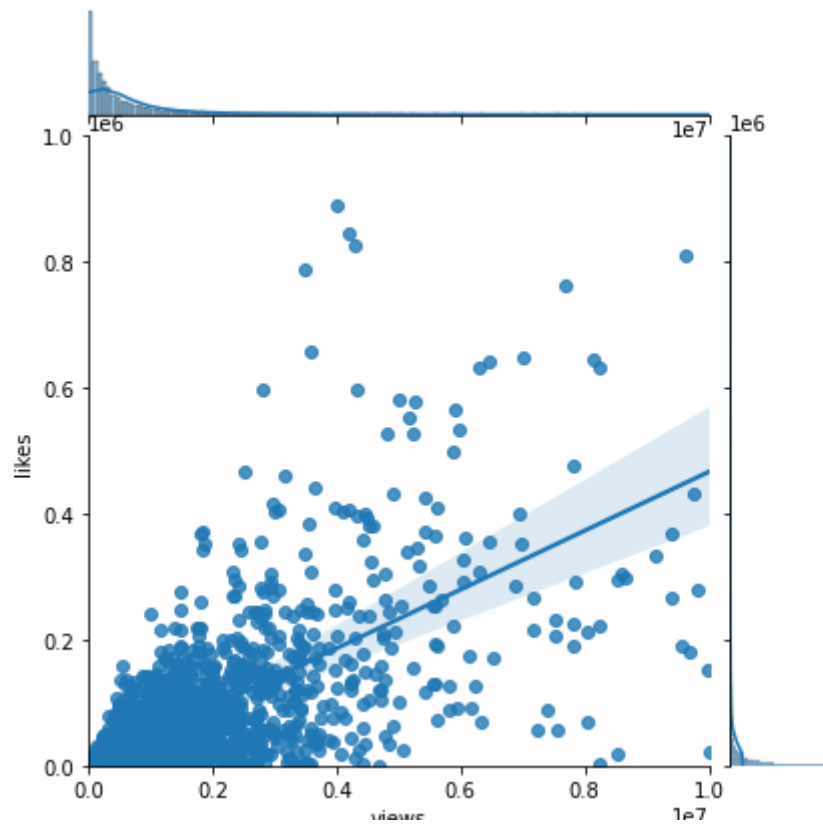
9. Топ каналов, чаще всего попадавших в тренды



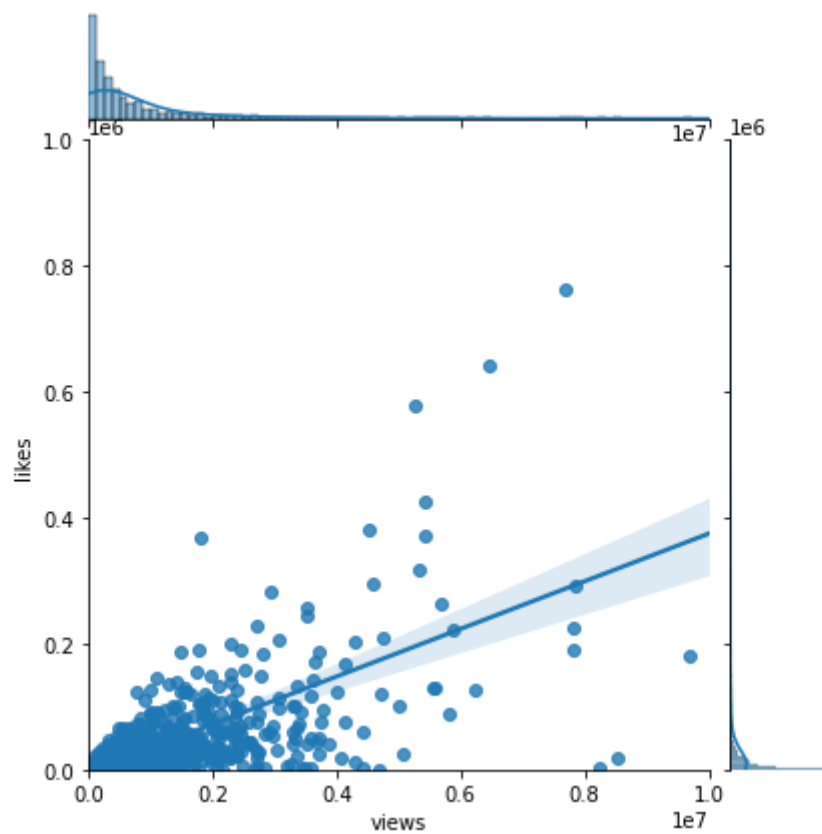
10. Распределение длин описания (в символах)



11. Отношение количества лайков к количеству просмотров



12. Отношение количества лайков к количеству просмотров для видео категории Entertainment (id = 24)



Сравнивая первые графики с полученными выше, получены ожидаемые результаты.

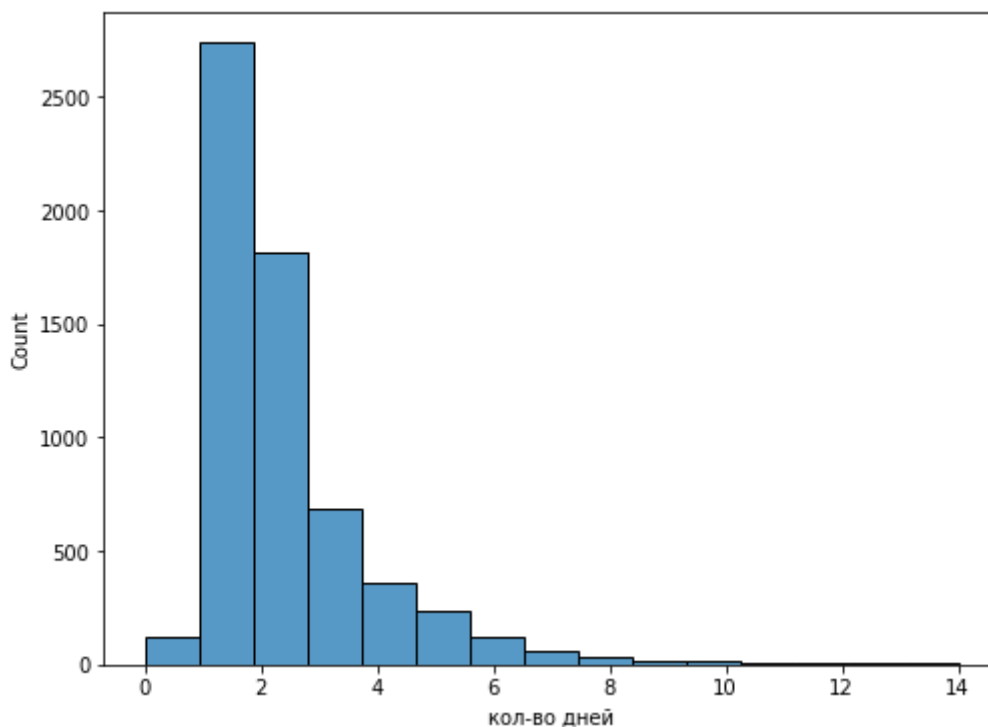
Теперь проверим гипотезы.

1. Гипотеза: видео попадают в тренды в среднем за 3 дня

```
]# гипотеза 1: видео попадают в тренды в среднем за 3 дня
days_to_trends = df.drop_duplicates(subset="video_id", keep="first")[["trending_date", "publish_time"]]
days_to_trends["publish_time"] = days_to_trends["publish_time"].dt.floor(freq="D")
delta = days_to_trends["trending_date"] - days_to_trends["publish_time"]
print(np.mean(delta))
print(np.max(delta))
sns.histplot([d.days for d in delta if d.days < 15], bins=15)
plt.xlabel("кол-во дней")
plt.show()
# гипотеза опровергнута - среднее значение составляет 22.5 дня
# некоторые видео попадают в тренды спустя более 10 лет после добавления на youtube
```

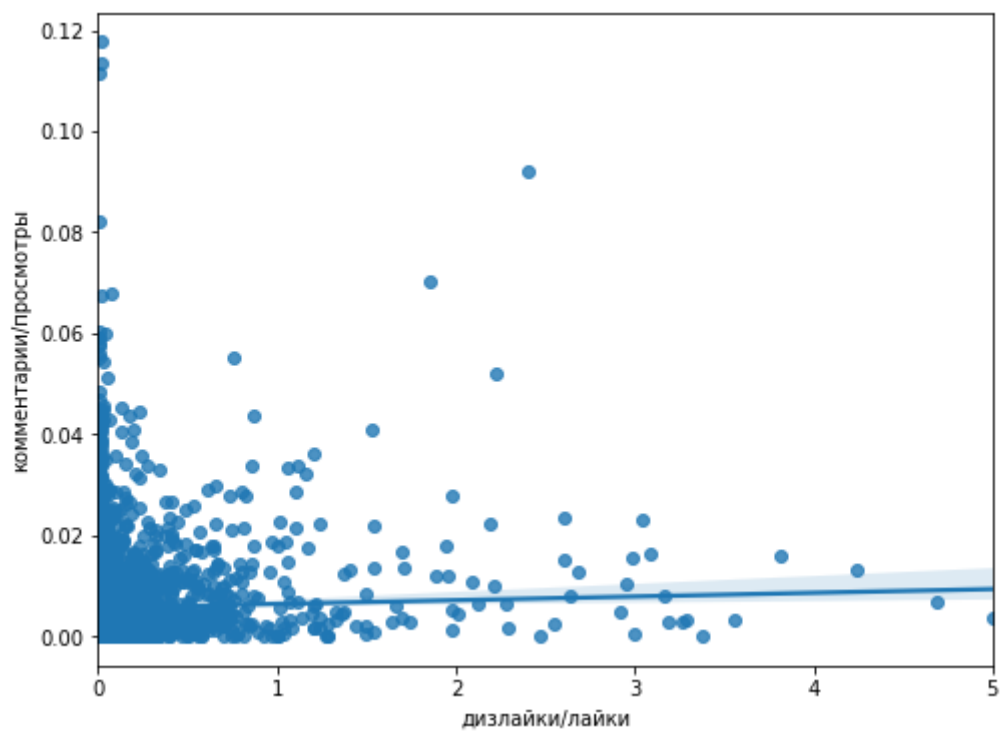
```
22 days 12:29:35.342465
4215 days 00:00:00
```

Гипотеза опровергнута - среднее значение составляет 22.5 дня, некоторые видео попадают в тренды спустя более 10 лет после добавления на youtube. Но если исключить влияние старых видеороликов, и ограничить только 15 днями.



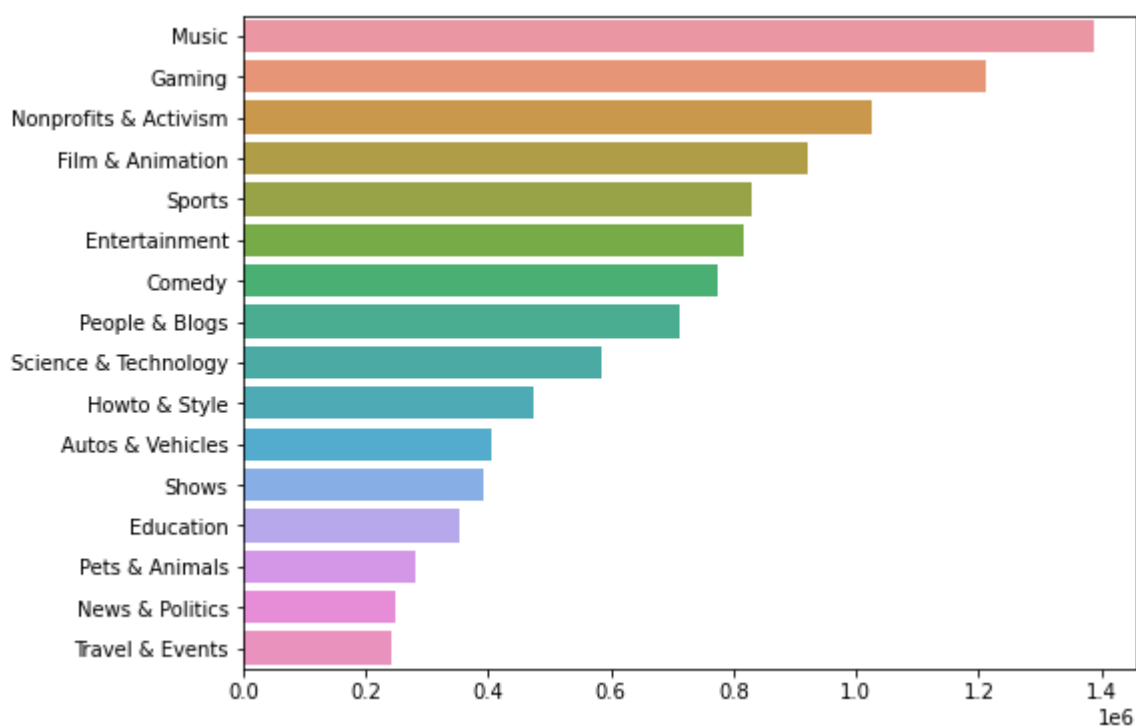
То действительно, чаще всего видеоролик попадает в тренды за 2-3 дня.

2. Гипотеза: при большем отношении дизлайков к лайкам выше отношение комментариев к просмотрам (когда зрителю не нравится видео он пойдет писать гневный комментарий).



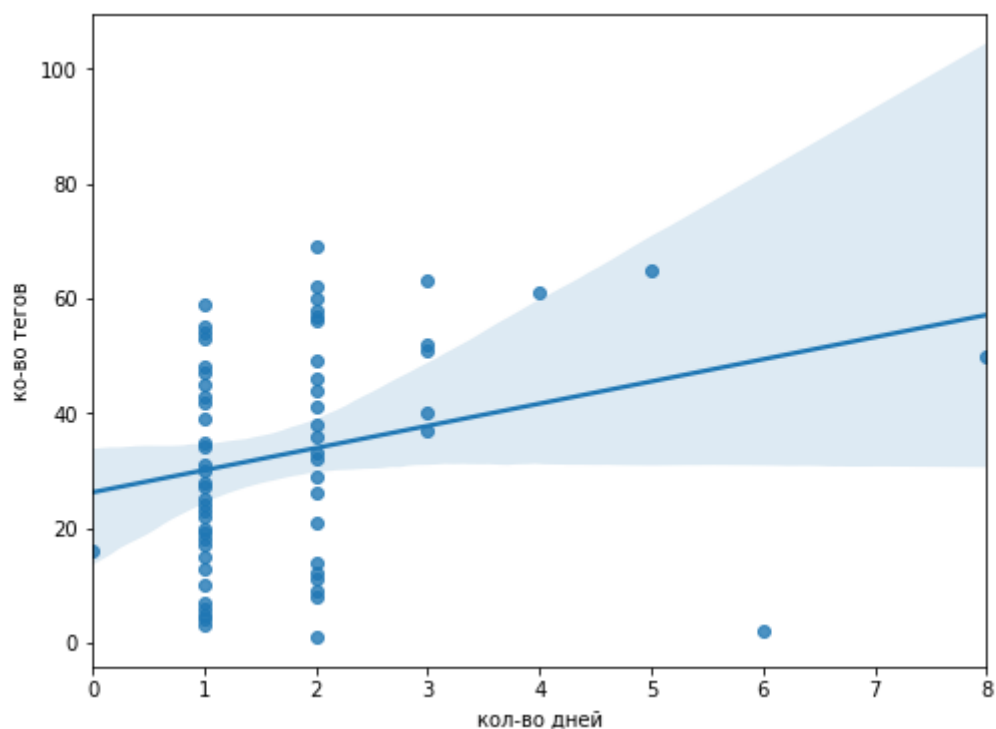
гипотеза подтверждена - по линии тренда видим небольшую прямую зависимость

3. Гипотеза: из типов видео больше всего просмотров в среднем набирают клипы



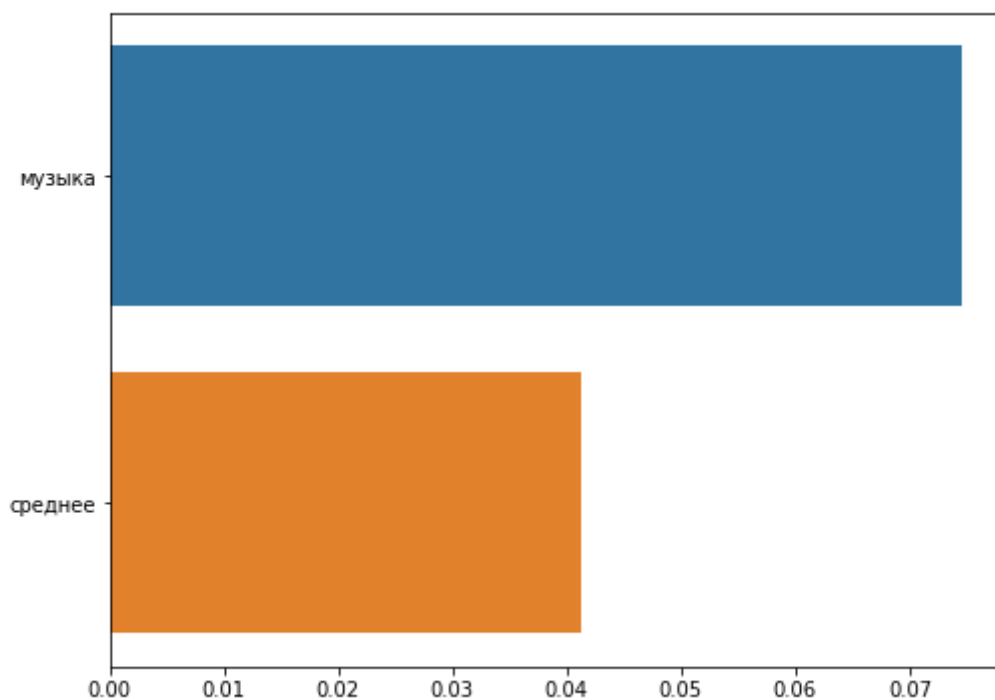
Гипотеза подтверждена, музыкальные клипы имеют отрыв от следующих категорий

4. Гипотеза: видео с большим количеством тегов быстрее попадает в тренды



Гипотеза отпровергнута - при большем количестве тегов попадание в тренды дольше. Если судить по линейной зависимости выше.

5. Гипотеза: на клипах меньше отношение лайков к просмотрам, чем у других видео в среднем (так как их просмотры сильно раздуты от пересматривания)



Гипотеза опровергнута. Число лайков на видеороликах в среднем сильно больше, чем в среднем.

Вывода: Проведена вновь описательная статистика датасета, проанализированны гипотезы предыдущие гипотезы. А также проделанна работа по изучению seaborn.